# Flavor Tagging Using Graph Neural Networks

Constructing and Calibrating GNNs for Flavor Tagging Jets in ATLAS and CMS

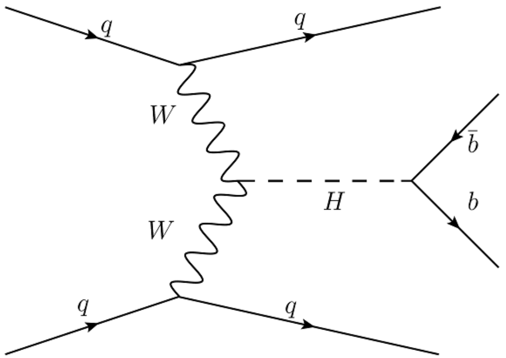
*Luke Vaughan*
*Oklahoma State University*

# Introduction

Today's discussion will include:

- Motivation for b-tagging
- Machine Learning in HEP
- Data Collection and Jet Reconstruction
- Deeps Sets and Message Passing Neural Networks
- Current GNN Implementations in ATLAS and CMS
- Calibration of Taggers
- Unfolding

# Why b-tagging?



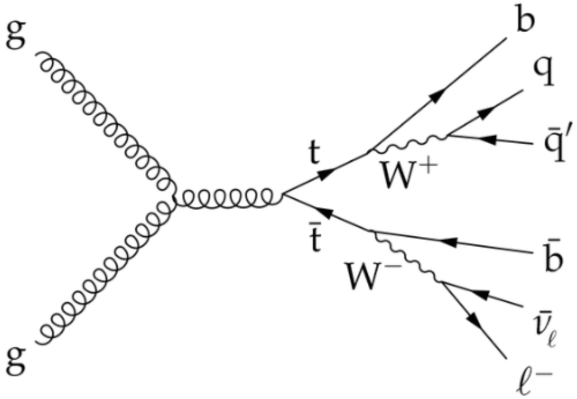




**Table 11.3:** The branching ratios and the relative uncertainty [44, 45] for a SM Higgs boson with $m_H = 125$ GeV.

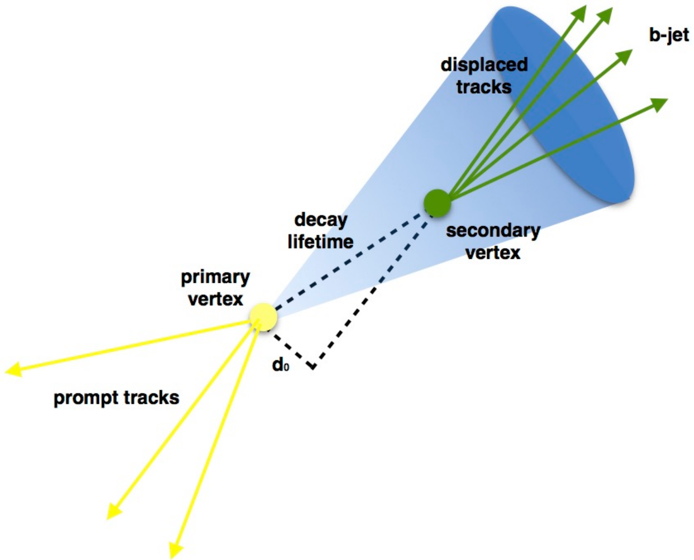
| Decay channel | Branching ratio | Rel. uncertainty |
|---|---|---|
| $H \to \gamma\gamma$ | $2.27 \times 10^{-3}$ | $^{+5.0\%}_{-4.9\%}$ |
| $H \to ZZ$ | $2.62 \times 10^{-2}$ | $^{+4.3\%}_{-4.1\%}$ |
| $H \to W^+W^-$ | $2.14 \times 10^{-1}$ | $^{+4.3\%}_{-4.2\%}$ |
| $H \to \tau^+\tau^-$ | $6.27 \times 10^{-2}$ | $^{+5.7\%}_{-5.7\%}$ |
| $H \to b\bar{b}$ | $5.84 \times 10^{-1}$ | $^{+3.2\%}_{-3.3\%}$ |
| $H \to Z\gamma$ | $1.53 \times 10^{-3}$ | $^{+9.0\%}_{-8.9\%}$ |
| $H \to \mu^+\mu^-$ | $2.18 \times 10^{-4}$ | $^{+6.0\%}_{-5.9\%}$ |

B-tagging is important for any physics process that includes b-jets in their final state.

Most notably, both Higgs Boson and Top Quark have large branching ratios to b quark, which demands that experimentalist focus their attention on b-tagging.

Source: PDG

# B-Hadron Unique Properties




B-Hadrons have many distinct properties such as long lifetime, high number of tracks produced in its decay, as well as displaced secondary vertex compared to primary vertex.

Typically, a B-hadron will travel several millimeters before decaying which causes the formation of a displaced secondary vertex.

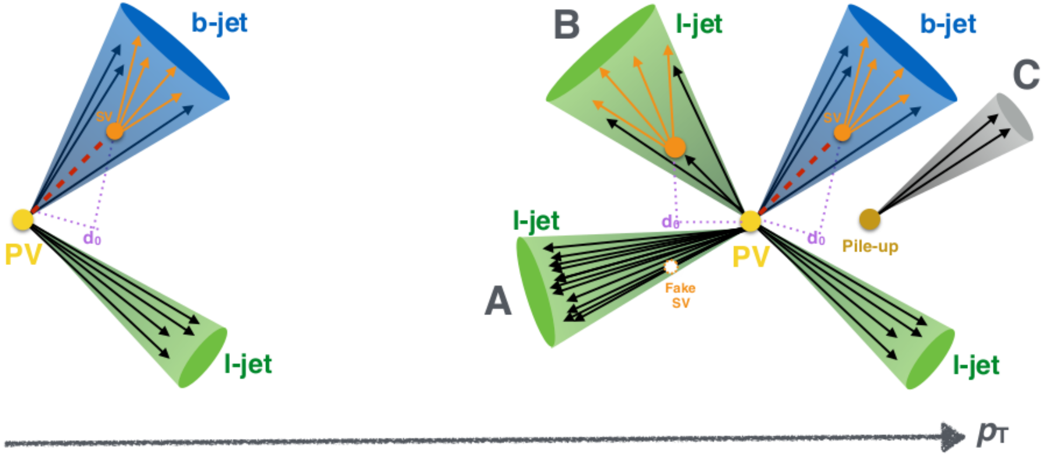
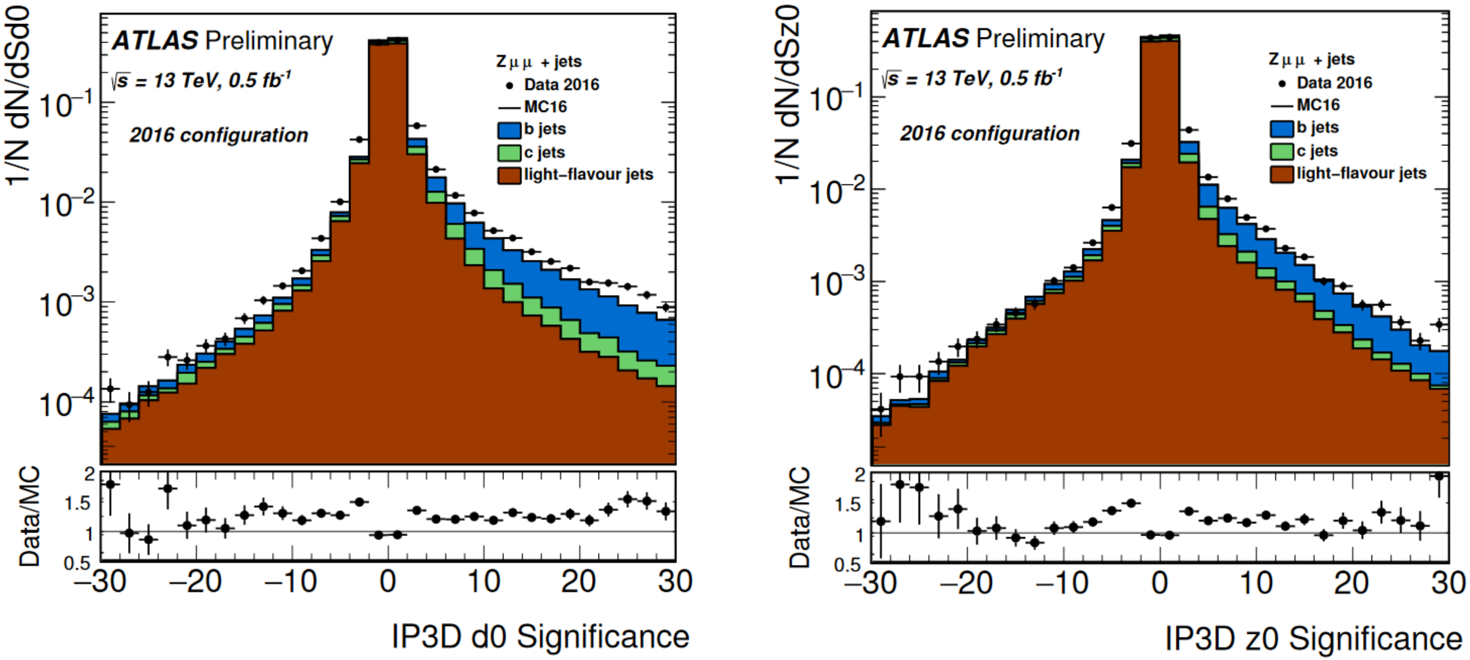



Image Credit: here

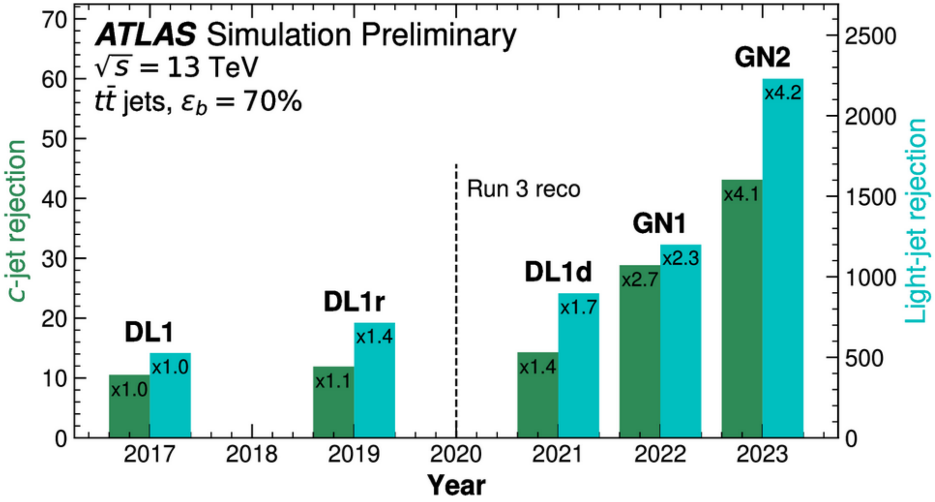
# B-Hadron Impact Parameters




As can be seen in the histograms above, b jets show significant separation for transverse, d0, and longitudinal, z0, impact parameter significance.

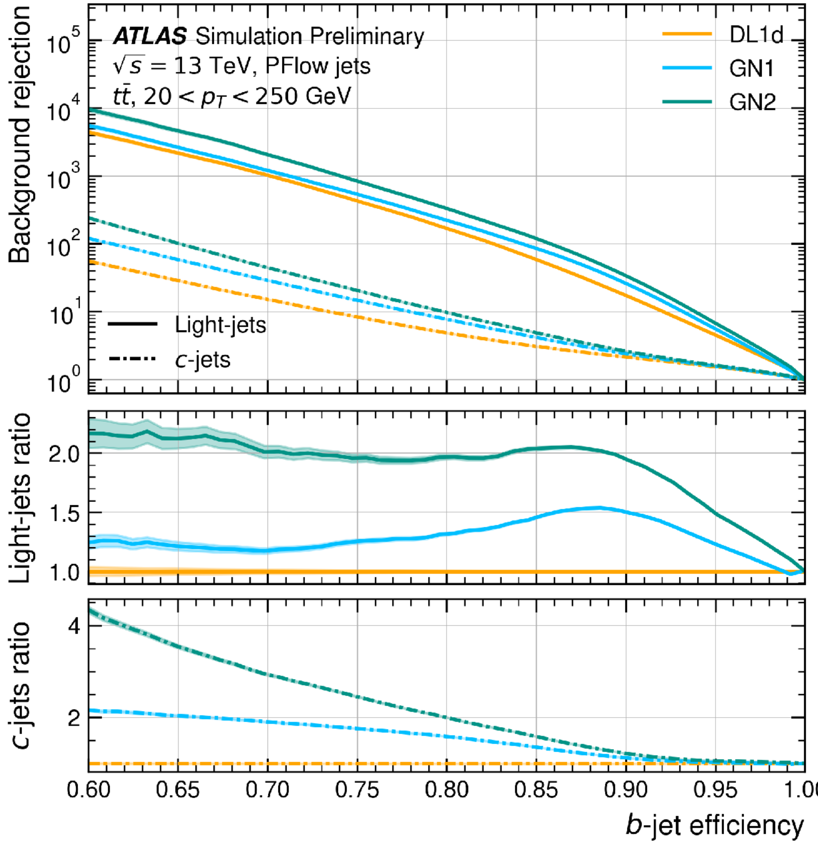
Histograms source: ATL-PHYS-PUB-2017-013
Secondary Vertex Algorithm: ATL-PHYS-PUB-2017-011

# Effective B-Tagging Methods




To maximize b-jet efficiency vs background rejection, it is useful to explore ML methods.

History of B-Tagging: Slide 14 here.




Source: ATLAS FTAG Public Plots

# Machine Learning in HEP

HEP currently uses ML in nearly all aspects of experimental, phenomenological, and theoretical analyses.

ML is currently used in b-tagging, tau ID, search for new physics, hardware quality control, anomaly detection, unfolding, and many others.

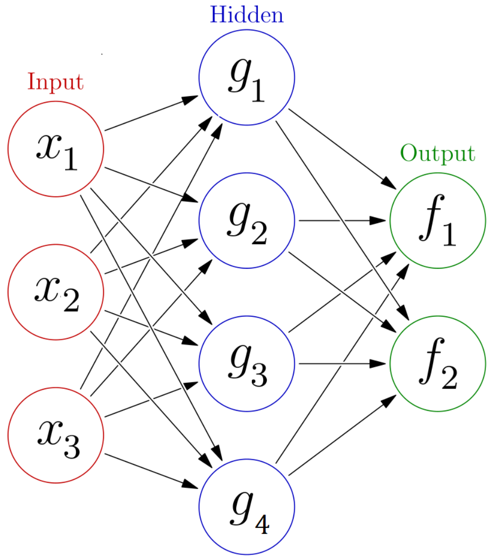
A comprehensive list of current uses for ML in HEP are shown in the link below.

Useful Link: HEPML-LivingReview

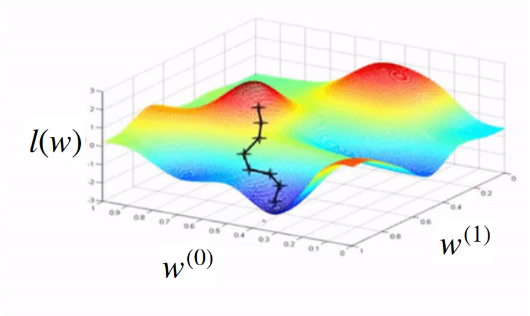

Image Credit: Javier Duarte

# Simple Neural Networks

A simple neural network is a directed acyclic graph that consists of neurons which use a weighted sum over the inputs of the previous layer and an activation function, K, to generate an output.




$$f(x) = K\left(\sum_i w_i g_i(x)\right)$$




The weights are "learned" (or rather optimized) during the training process by performing gradient descent of the loss function by evaluating the loss for predicted labels vs true labels. Common loss functions include cross entropy for classification or MSE for regression.

Source: Mathematics of NNs
Credit: LBNL ML July 2023 Workshop

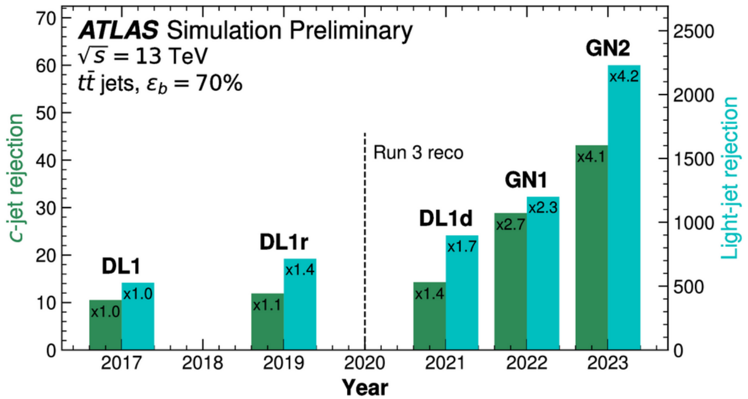
# Universal Approximation Theorem

**Universal approximation theorem** — Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset $X$ of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes $\sigma$ applied to each component of $x$.

Then $\sigma$ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m)$, $\varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m \times k}$ such that

$$\sup_{x \in K} \|f(x) - g(x)\| < \varepsilon$$

where $g(x) = C \cdot (\sigma \circ (A \cdot x + b))$




Arbitrarily deep and wide neural networks allow us to approximate any function. If deep NN can approximate any function, why have we seen performance improvement by switching to GNN? The answer lies in better data representation.

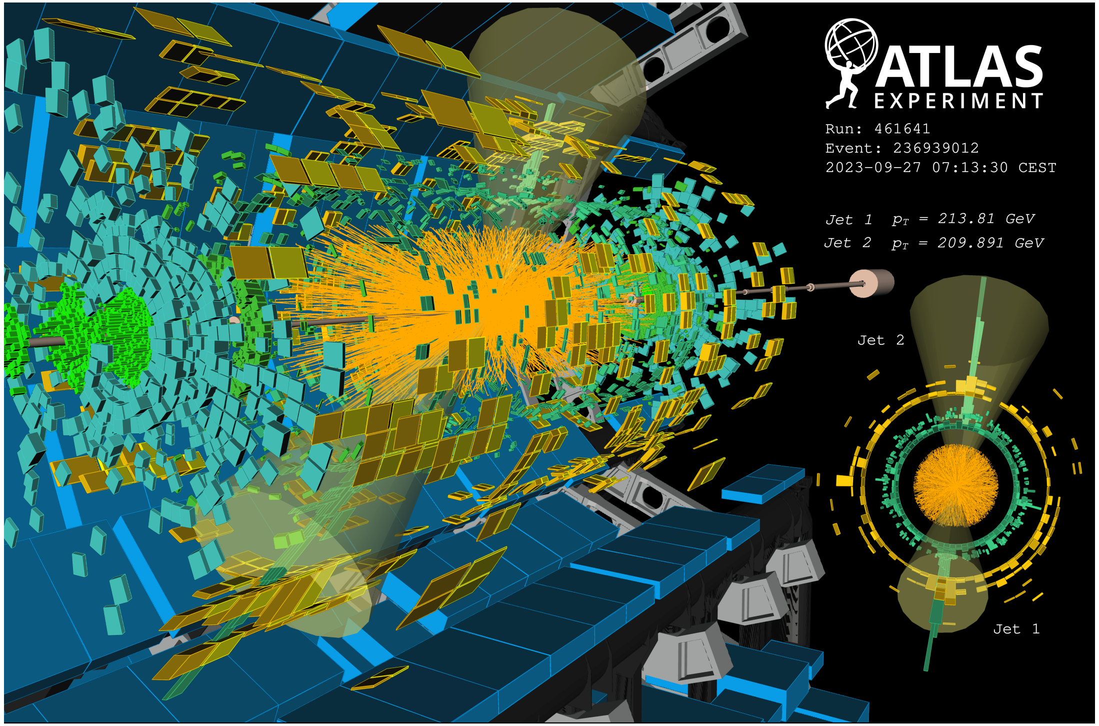
Source: Universal Approximation Theorem Wikipedia

# Data Collection at ATLAS

Typical events in the ATLAS detector are very "busy". What variables/objects would be most useful for the NN?

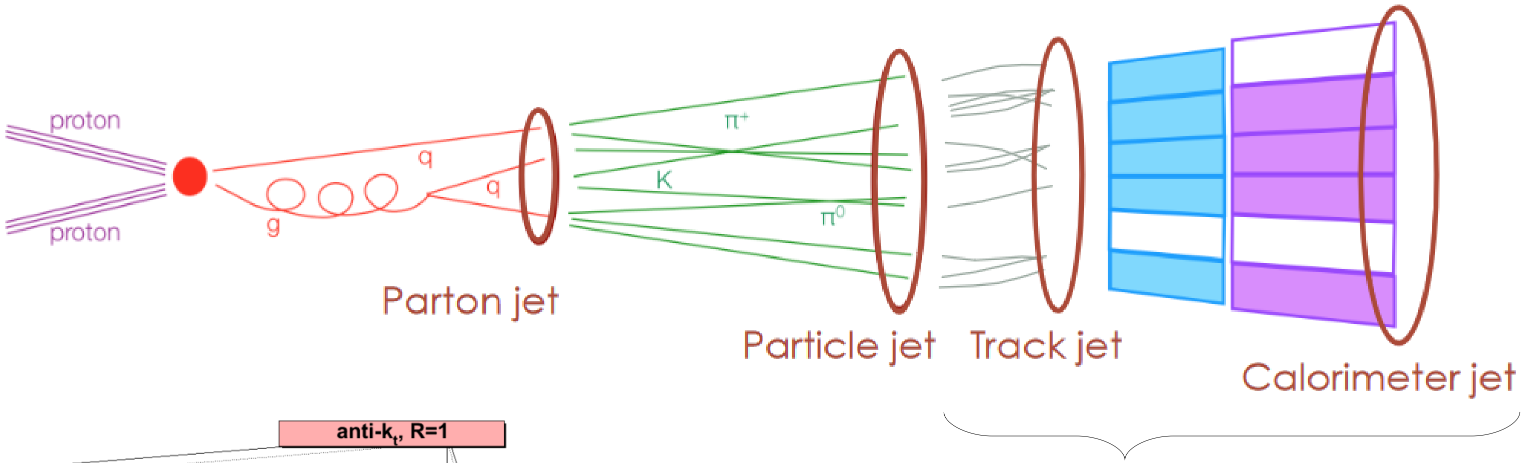
Data is collected as readout from sensors from two main components: the tracker and the calorimeter.

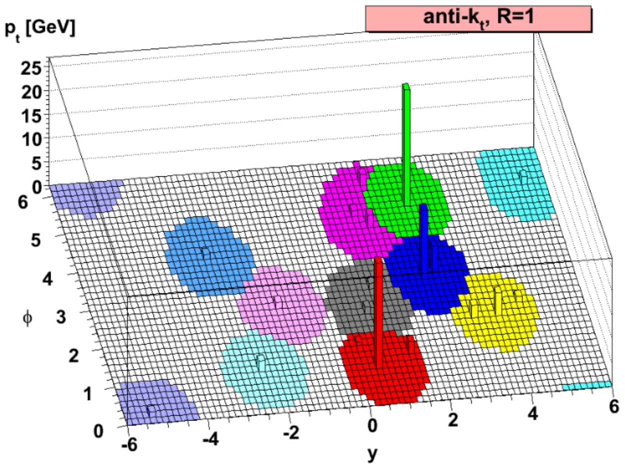
How do we *reconstuct* the readout from these sensors as tracks and jets? How do we best represent this data for a NN?




Source: ATLAS Public Event Dislay
Source: Event Display Public Results

# Jet Reconstruction in ATLAS




Current reconstruction techniques at ATLAS use tracker hits and Calorimeter deposits to reconstruct PFlow Jets.

The anti-kt algorithm, shown left, is used to cluster jets.

Credit: ATLAS Hadronic Calibration Workshop
Anti-$k_t$ Source: arxiv 0802.1189

# Low Level Inputs

Table 2: Input features to the GN1 model. Basic jet kinematics, along with information about the reconstructed track parameters and constituent hits are used. Shared hits, are hits used on multiple tracks which have not been classified as split by the cluster-splitting neural networks [15], while split hits are hits used on multiple tracks which have been identified as merged. A hole is a missing hit, where one is expected, on a layer between two other hits on a track. The track leptonID is an additional input to the GN1 Lep model.

| Jet Input | Description |
|---|---|
| $p_T$ | Jet transverse momentum |
| $\eta$ | Signed jet pseudorapidity |

| Track Input | Description |
|---|---|
| $q/p$ | Track charge divided by momentum (measure of curvature) |
| $\mathrm{d}\eta$ | Pseudorapidity of the track, relative to the jet $\eta$ |
| $\mathrm{d}\phi$ | Azimuthal angle of the track, relative to the jet $\phi$ |
| $d_0$ | Closest distance from the track to the PV in the longitudinal plane |
| $z_0 \sin\theta$ | Closest distance from the track to the PV in the transverse plane |
| $\sigma(q/p)$ | Uncertainty on $q/p$ |
| $\sigma(\theta)$ | Uncertainty on track polar angle $\theta$ |
| $\sigma(\phi)$ | Uncertainty on track azimuthal angle $\phi$ |
| $s(d_0)$ | Lifetime signed transverse IP significance |
| $s(z_0)$ | Lifetime signed longitudinal IP significance |
| nPixHits | Number of pixel hits |
| nSCTHits | Number of SCT hits |
| nIBLHits | Number of IBL hits |
| nBLHits | Number of B-layer hits |
| nIBLShared | Number of shared IBL hits |
| nIBLSplit | Number of split IBL hits |
| nPixShared | Number of shared pixel hits |
| nPixSplit | Number of split pixel hits |
| nSCTShared | Number of shared SCT hits |
| nPixHoles | Number of pixel holes |
| nSCTHoles | Number of SCT holes |
| leptonID | Indicates if track was used in the reconstruction of an electron or muon (only for GN1 Lep) |

Once the event has been reconstructed, we have access to all the information shown in the table.

Calorimeter information includes pT, eta, and phi of the jet.

Tracker information includes kinematics, secondary vertex information, and tracker hits.

Source: ATL-PHYS-PUB-2022-027

# Deep Set Data Representation

For the purpose of flavor tagging, each jet is represented by an unordered collection of tracks. The number of tracks per jet can vary and the order of the tracks should not affect the output of the classifier.

How do we learn from set data? Deep sets.

**Theorem 2** *A function $f(X)$ operating on a set $X$ having elements from a countable universe, is a valid set function,* i.e., ***invariant*** *to the permutation of instances in $X$, iff it can be decomposed in the form $\rho\left(\sum_{x \in X} \phi(x)\right)$, for suitable transformations $\phi$ and $\rho$.*

*The foundation of Message Passing GNNs*

Note: Deep sets representation is similar to point cloud representation.

Source: arxiv 1703.06114

# Message Passing Graph Neural Networks

$$\mathbf{x}_i^{(k)} = \overset{3}{\gamma^{(k)}} \left( \mathbf{x}_i^{(k-1)}, \overset{2}{\bigoplus_{j \in \mathcal{N}(i)}} \overset{1}{\phi^{(k)}} \left( \underbrace{\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}, \mathbf{e}_{j,i}}_{\text{input features: self, neighbor, and edge}} \right) \right)$$

1. Prepare Message
   - Use NN that learns the optimal message
2. Aggregate Messages
   - Use a permutation invariant operation on the messages such as sum, mean, or max.
3. Update Node Embeddings
   - Use NN that learns the optimal node embedding update.

Source: PyTorch Documentation

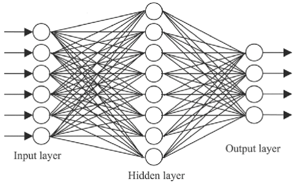
# Message Passing Graph Neural Networks

*Consider a jet composed of a set of tracks:* $J = \{t_0, t_1, t_2, \ldots, t_n\}$

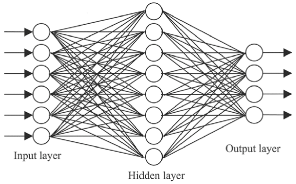
<u>Neighboring Tracks</u>

1

$t_1 = \{pT, \eta, \phi\}$

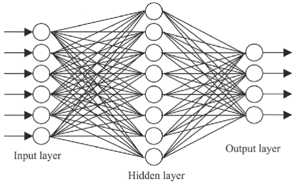



$t_2 = \{pT, \eta, \phi\}$




$\vdots$

$t_n = \{pT, \eta, \phi\}$



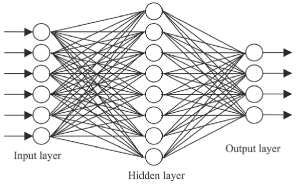

2

<u>Self Track +</u>
<u>Aggregated Message</u>

$t_0 = \{pT, \eta, \phi\}$

$\oplus$

3




$t_0 = \{f_1, f_2, f_3, \ldots\}$

1. Prepare Message
2. Aggregate Messages
3. Update Node Embeddings

# ATLAS GN1

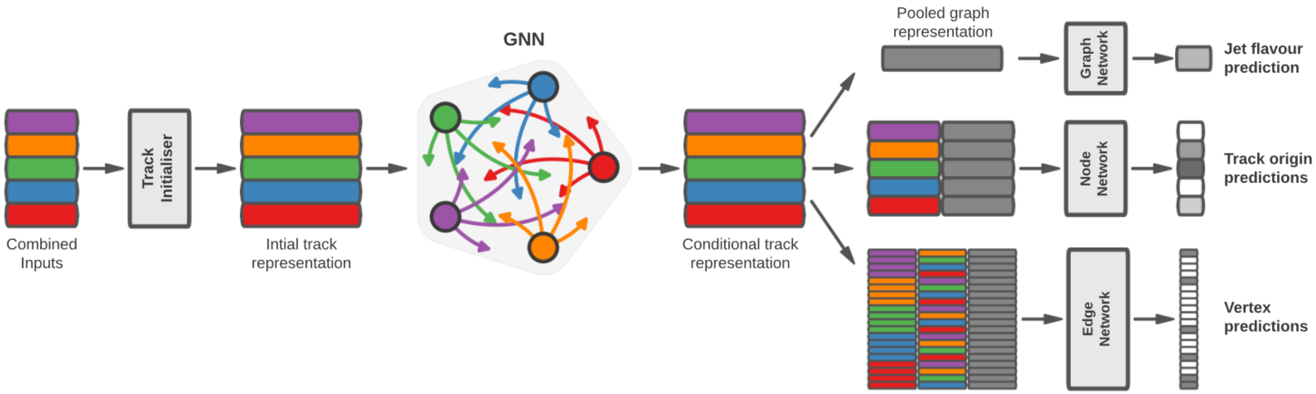



Figure 3: The network architecture of GN1. Inputs are fed into a per-track initialisation network, which outputs an initial latent representation of each track. These representations are then used to populate the node features of a fully connected graph network. After the graph network, the resulting node representations are used to predict the jet flavour, the track origins, and the track-pair vertex compatibility.

The GNN in ATLAS uses message passing neural network to update the node representation of each track based on its neighbors. Then jet classification can be done using a global graph network.

Source: ATL-PHYS-PUB-2022-027

# ATLAS GN1

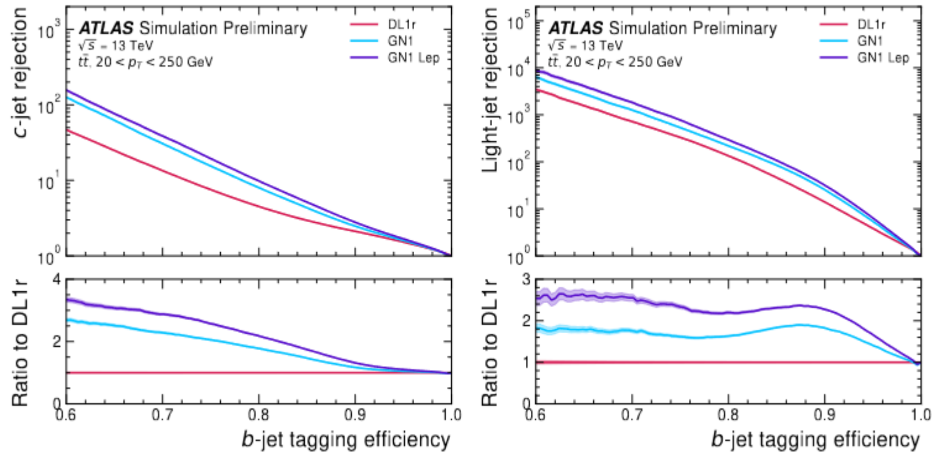



Figure 5: The $c$-jet (left) and light-jet (right) rejections as a function of the $b$-jet tagging efficiency for jets in the $t\bar{t}$ sample with $20 < p_{\mathrm{T}} < 250$ GeV. The ratio with respect to the performance of the DL1r algorithm is shown in the bottom panels. A value of $f_c = 0.018$ is used in the calculation of $D_b$ for DL1r and $f_c = 0.05$ is used for GN1 and GN1 Lep. Binomial error bands are denoted by the shaded regions. At $b$-jet tagging efficiencies less than ~75%, the light-jet rejection becomes so large that the effect of the low number of jets is visible. The lower $x$-axis range is chosen to display the $b$-jet tagging efficiencies usually probed in these regions of phase space.

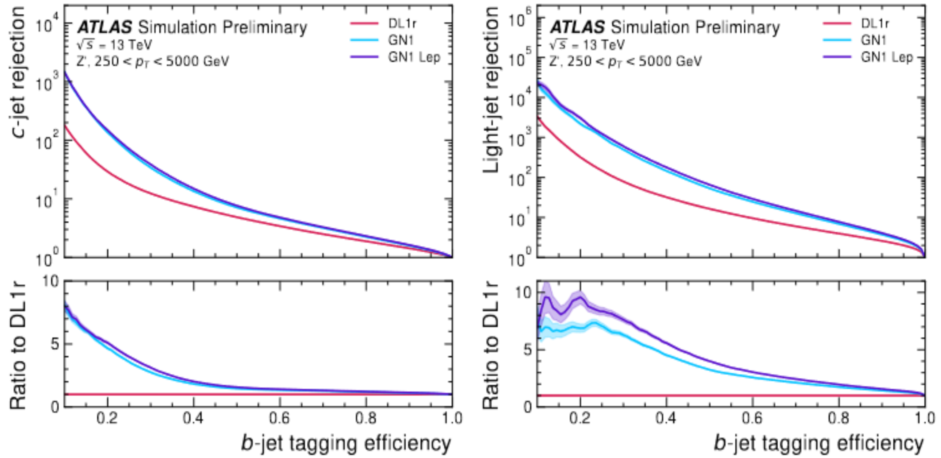
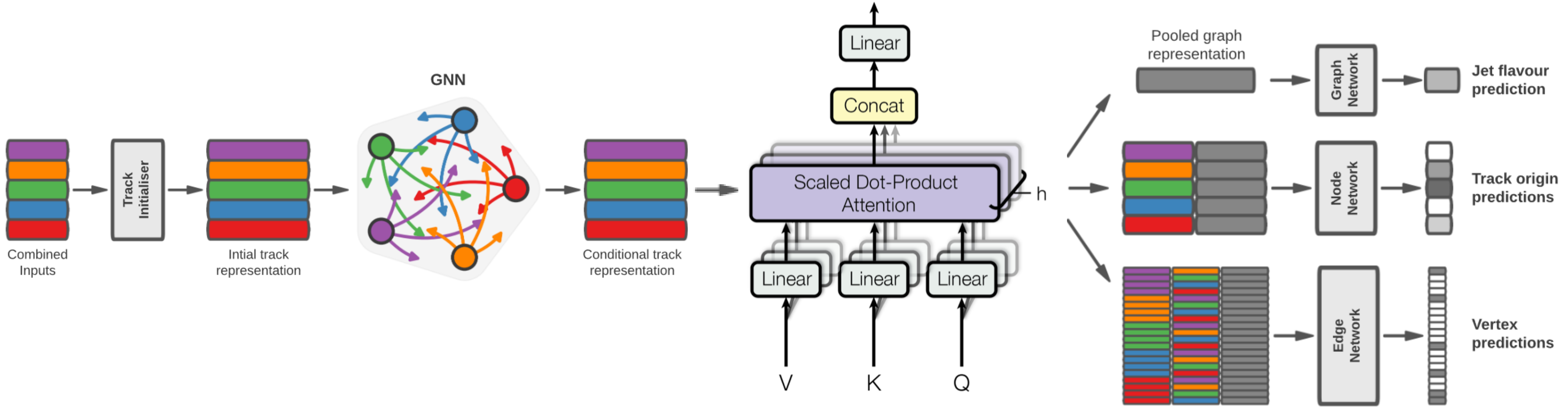



Figure 6: The $c$-jet (left) and light-jet (right) rejections as a function of the $b$-jet tagging efficiency for jets in the $Z'$ sample with $250 < p_{\mathrm{T}} < 5000$ GeV. The ratio with respect to the performance of the DL1r algorithm is shown in the bottom panels. A value of $f_c = 0.018$ is used in the calculation of $D_b$ for DL1r and $f_c = 0.05$ is used for GN1 and GN1 Lep. Binomial error bands are denoted by the shaded regions. At $b$-jet tagging efficiencies less than ~20%, the light-jet rejection becomes so large that the effect of the low number of jets is visible. The lower $x$-axis range is chosen to display the $b$-jet tagging efficiencies usually probed in these regions of phase space.

Previous model used in ATLAS, DL1r, uses Deep Learning approach. Notice there is a large improvement in c-jet and light-jet rejection by using a deep set data representation.

Source: ATL-PHYS-PUB-2022-027

# ATLAS GN2




GN2 uses the same architecture as GN1 but includes a multi-head attention layer after the updated GNN node embeddings. This has been shown to increase performance.

GN2X: ATL-PHYS-PUB-2023-021

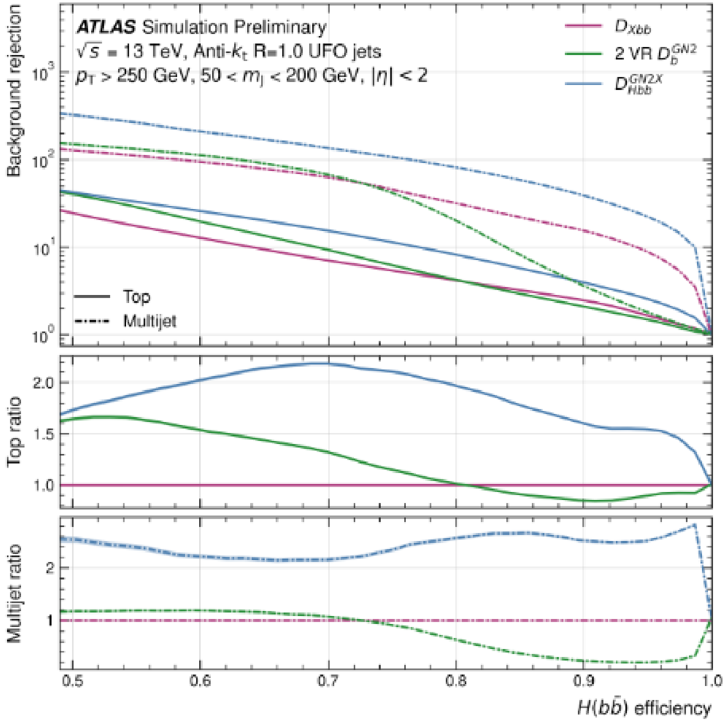
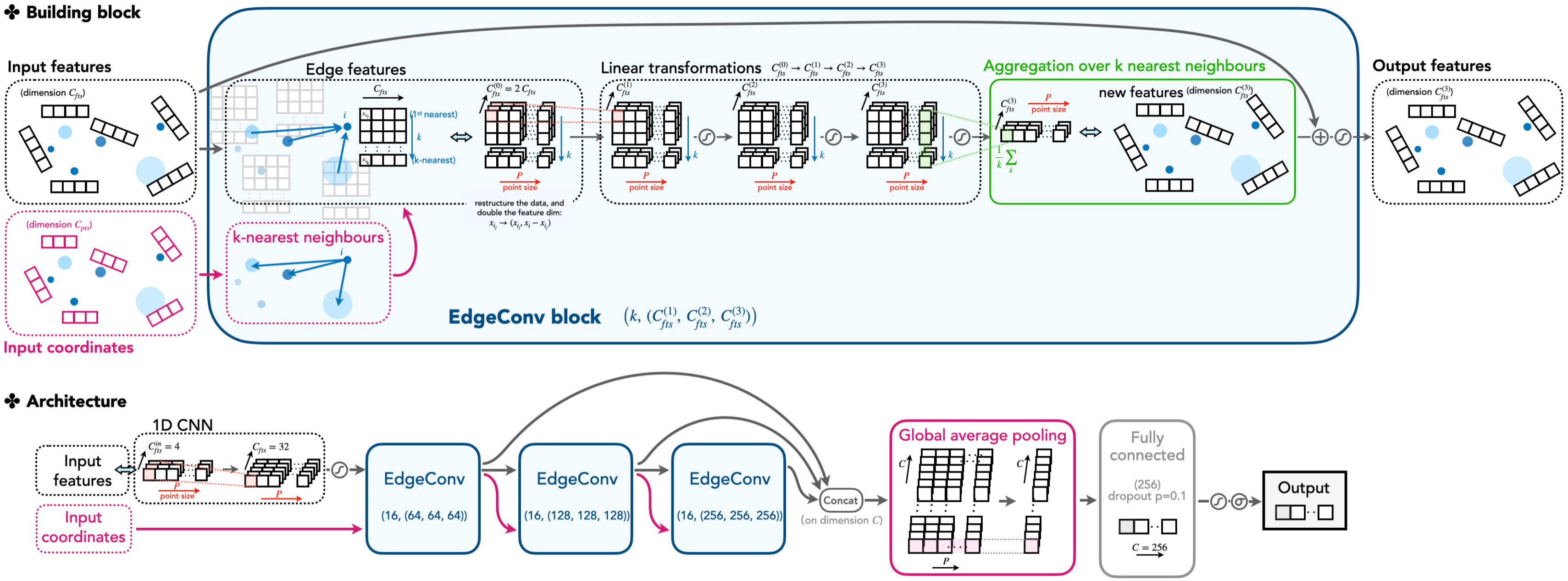



Figure 2: Top and multijet rejections as a function of the $H(b\bar{b})$ efficiency for jets with $p_T > 250$ GeV and mass $50\,\text{GeV} < m_J < 200\,\text{GeV}$. Performance of the GN2X algorithm is compared to the $D_{Xbb}$ and VR subjets baselines. Statistical uncertainty bands (calculated with a binomial model) are denoted. The distribution is shown for the SM evaluation samples.

GN2X is a variant of GN2 and is used to classify boosted jets as H→bb, H→cc, top, or multijet (QCD).

GN2X along with GN2 will be crucial for analyses with hadronic channels in LHC Run 3.

GN2X: ATL-PHYS-PUB-2023-021




*Full Schematic of CMS ParticleNet*

ParticleNet and GN1 are both examples of message passing neural networks. However, ParticleNet uses a clever trick by performing dynamic edge convolution using k-nearest neighbors which allows ParticleNet to *learn connections between subjets*.

Source: ParticleNet Documentation

# CMS ParticleNet

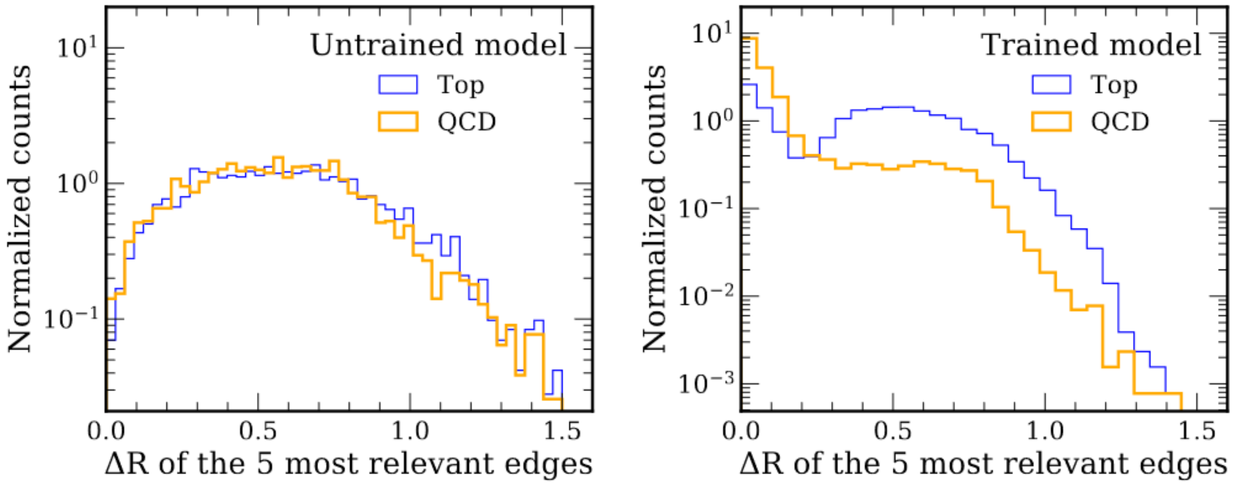



Figure 2: We present the distribution of the $\Delta R$ of the five most relevant edges for top quark jets (blue) versus QCD jets (orange) for an untrained ParticleNet model (left) and the learned distribution by a trained ParticleNet model (right). The result for the untrained model is an average over 10 randomly initialized models.

As ParticleNet performs dynamic edge convolutions, the distribution of the most relevant edges is affected differently for Top and QCD jets.

Source: arxiv 2211.09912

# CMS ParticleNet

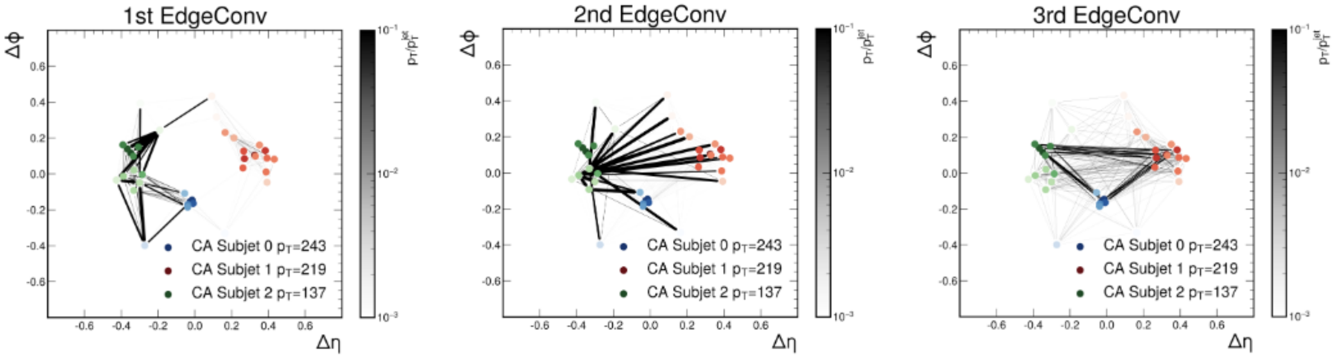
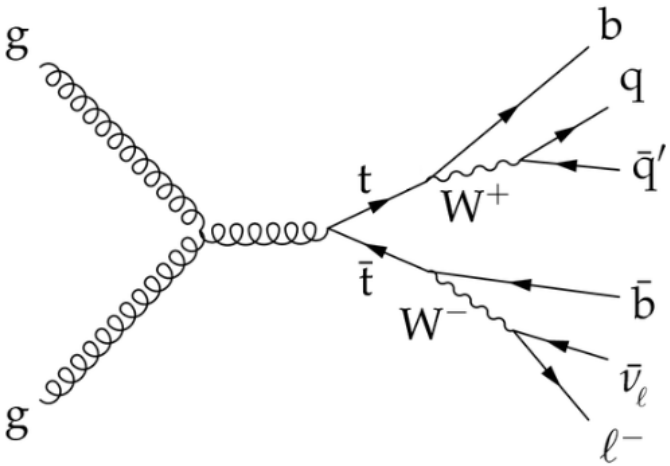


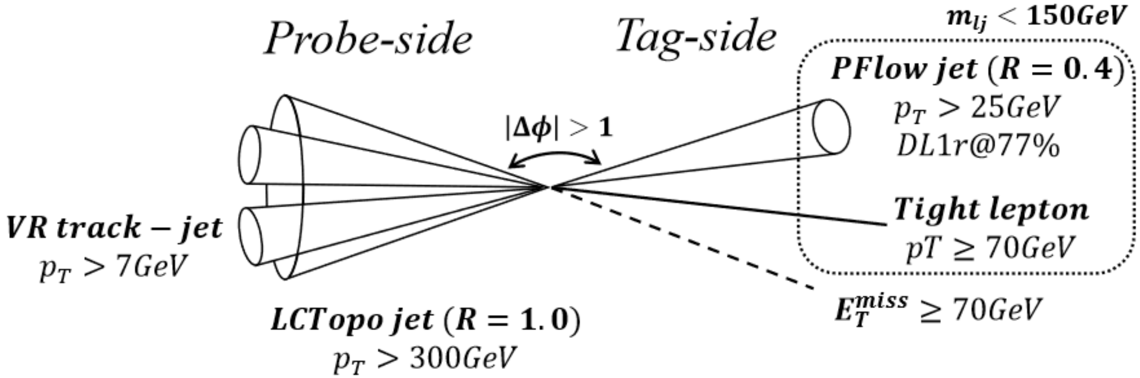

Figure 3: The three edge $R$ graphs for a true top quark jet corresponding to the three graphs learned with each EdgeConv block. The jet constituents are represented as nodes in $(\eta, \phi)$ space with interconnections as edges, whose intensities correspond to the connection's edge $R$ score. Each node's intensity corresponds to the relative $p_\mathrm{T}$ of the corresponding particle. Constituents belonging to the three different CA subjets are shown in blue, red, and green in descending $p_\mathrm{T}$ order. We observe that by the last EdgeConv block the model learns to rely more on edge connections between the different subjets.

As shown in the linked paper, the use of dynamic edge convolutions allows particle net to learn connections between subjets which is hinting at the physical substructure of the jet. QCD jets do not exhibit this behavior with ParticleNet.

Source: arxiv 2211.09912

# Calibration

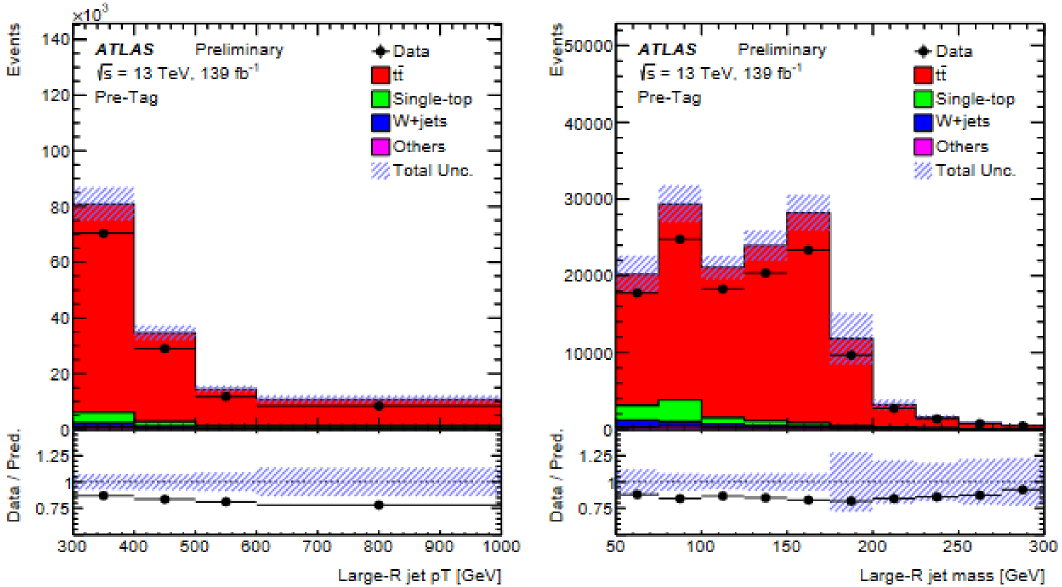



Semi-leptonic ttbar decay has a unique signature in the detector, and can be used for tagger calibration. The method is referred to as tag and probe method where the lepton, b-tagged jet, and MET are matched to a large-R jet.

Source: ATL-PHYS-PUB-2021-035

# Calibration




(a) (b)

Figure 9: Data and MC simulated distributions of $p_T$ (a) and mass of the probe jet (b) for the pre-tag selection. The ratio panel shows the data-to-MC ratio. The uncertainty band includes MC statistical and systematic uncertainties.

$$N_{tag}^{data} = \mu * \frac{\epsilon^{\text{data}}}{\epsilon^{\text{MC}}} * N_{tag}^{ttbar} + N_{tag}^{other}$$

$$N_{untag}^{data} = \mu * \frac{1-\epsilon^{\text{data}}}{1-\epsilon^{\text{MC}}} * N_{untag}^{ttbar} + N_{untag}^{other}$$

Extracting scale factors simply involves counting the number of events in the tagged (SR) and untagged (CR) and performing a likelihood fit.

Source: ATL-PHYS-PUB-2021-035

# Calibration

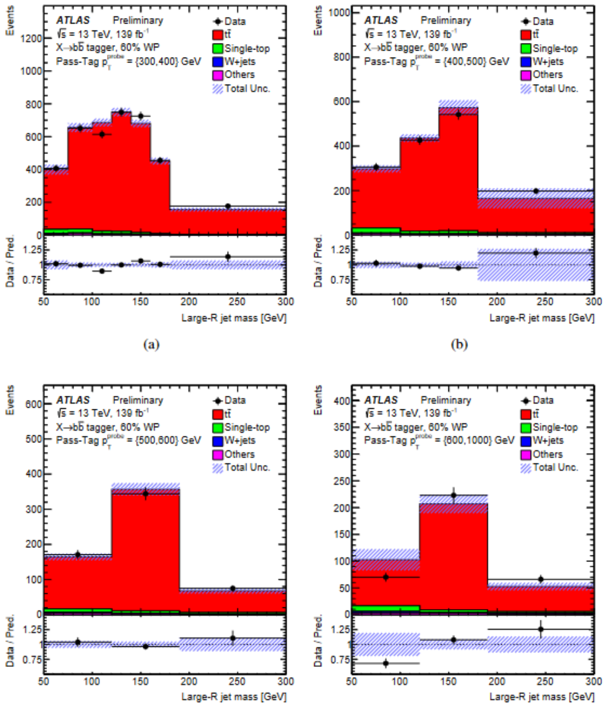



Figure 10: Post-fit distribution of the large-$R$ jet mass in $SR^{60\%}_{pass}$ used for calibration of the top jet mis-tag efficiency with $X \rightarrow b\bar{b}$ tagger at 60% efficiency WP for $300 < p_T^{probe} < 400$ GeV (a); $400 < p_T^{probe} < 500$ GeV (b); $500 < p_T^{probe} < 600$ GeV (c) and $600 < p_T^{probe} < 1000$ GeV (d). The uncertainty band represents the systematic uncertainty.

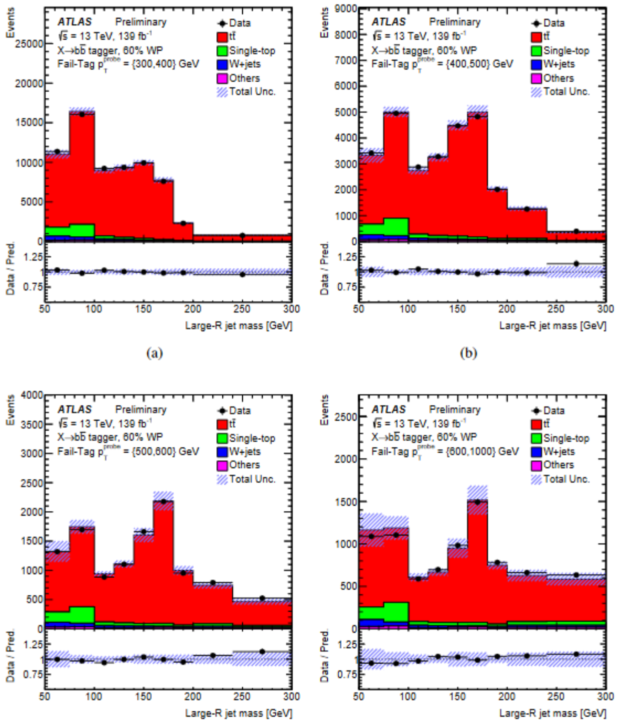
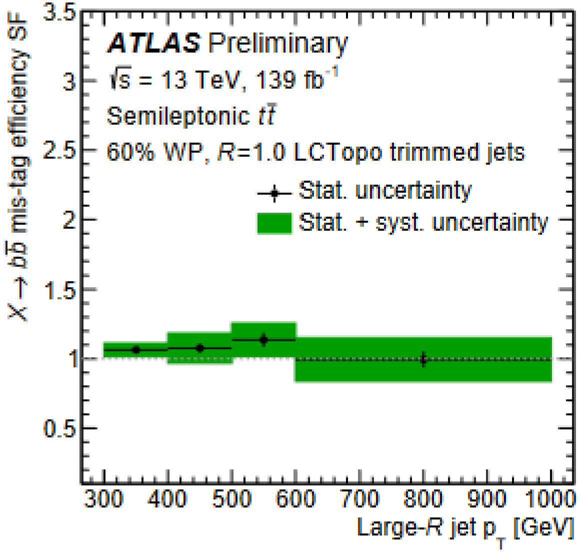



Figure 11: Post-fit distribution of the large-$R$ jet mass in $SR^{60\%}_{fail}$ used for calibration of the top jet mis-tag efficiency with $X \rightarrow b\bar{b}$ tagger at 60% efficiency WP for $300 < p_T^{probe} < 400$ GeV (a); $400 < p_T^{probe} < 500$ GeV (b); $500 < p_T^{probe} < 600$ GeV (c) and $600 < p_T^{probe} < 1000$ GeV (d). The uncertainty band represents the systematic uncertainty.

Post-fit distributions in SR, right, and CR, left.

Source: ATL-PHYS-PUB-2021-035

| $p_T$ [GeV] | 300 – 400 | 400 – 500 | 500 – 600 | 600 – 1000 |
|---|---|---|---|---|
| SF | 1.06 | 1.08 | 1.14 | 0.99 |
| Total unc. | 0.045 | 0.10 | 0.11 | 0.16 |
| Statistic unc. | 0.018 | 0.029 | 0.046 | 0.06 |
| Systematic unc. | 0.041 | 0.095 | 0.095 | 0.15 |
| $t\bar{t}$ modelling | 0.039 | 0.094 | 0.088 | 0.14 |
| $t\bar{t}$ PS | <0.001 | 0.002 | 0.003 | 0.002 |
| $t\bar{t}$ FSR | 0.022 | 0.075 | 0.036 | 0.093 |
| $t\bar{t}$ ISR | 0.031 | 0.055 | 0.078 | 0.11 |
| $t\bar{t}$ generator | <0.001 | <0.001 | <0.001 | <0.001 |
| $t\bar{t}$ PDF | 0.01 | 0.015 | 0.019 | 0.022 |
| $t\bar{t}$ cross-section | – | <0.001 | <0.001 | <0.001 |
| Single-top modelling | 0.007 | 0.009 | 0.020 | 0.023 |
| Single-top $Wt$ DR vs DS | 0.005 | 0.007 | 0.014 | 0.015 |
| Single-top PS | <0.001 | 0.002 | 0.007 | 0.015 |
| Single-top generator | 0.004 | – | 0.011 | 0.002 |
| Single-top cross-section | 0.003 | 0.002 | 0.003 | 0.003 |
| $W$ + jets ( scale, cross-section ) | 0.004 | 0.003 | 0.004 | 0.005 |
| Small-$R$ jet energy | 0.008 | 0.011 | 0.022 | 0.016 |
| Large-$R$ jet energy and mass | 0.004 | 0.008 | 0.014 | 0.008 |
| Small-$R$ jet Flavour tagging related | 0.001 | 0.001 | 0.001 | 0.002 |
| Others | 0.003 | 0.004 | 0.004 | 0.006 |

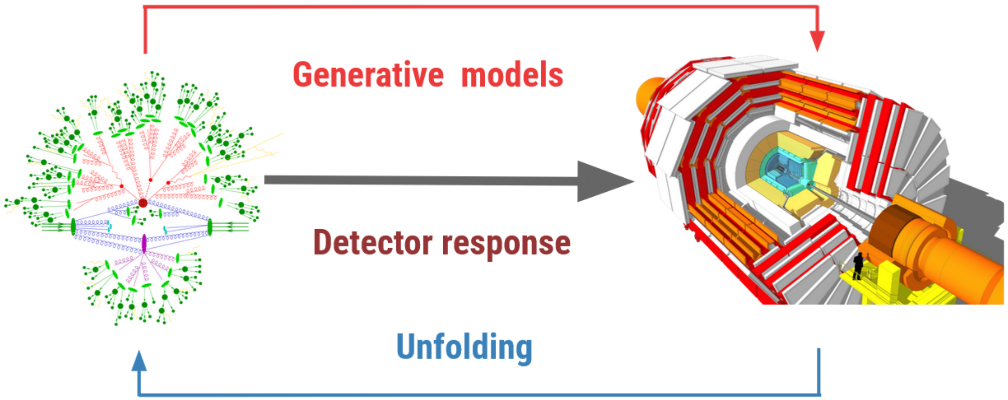



Breakdown of uncertainties on the scale factor are obtained with a non-profile liklihood fit.

Source: ATL-PHYS-PUB-2021-035

# Unfolding

Notice how this entire talk I have been referring to reconstruction level information. Reconstructed information has unwanted detector effects... (•`_´• )




What if we want to unfold the detector level measurement to get truth level information?

Credit: Returning CP-Observables to The Frames They Belong
Image Source: LBNL ML Workshop 2023 Unfolding Slides

# Conclusion

Thank you for listening! (•‿•)

Questions?

- Motivation for b-tagging
- Machine Learning in HEP
- Data Collection and Jet Reconstruction
- Deeps Sets and Message Passing Neural Networks
- Current GNN Implementations in ATLAS and CMS
- Calibration of Taggers
- Unfolding